Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Environmental Sciences 4 (2011) 50–55

Procedia Environmental Sciences

1st Conference on Spatial Statistics 2011: Mapping Global Change

Modeling species distribution dynamics with SpatioTemoral Exploratory Models: Discovering patterns and processes of broad-scale avian migrations

Daniel Fink, Wesley M. Hochack, Benjamin Zuckerberg, and Steve T. Kelling

Cornell Lab of Ornithology, 159 Sapsucker Woods Rd, Ithaca NY 14850, USA, df36@cornell.edu

Abstract

The distributions of animal populations are not static. During regular migratory movements species exploit different habitats. This spatiotemporal variation needs to be accounted for when modeling a species' distribution and is essential for developing conservation strategies for widespread species, and especially for migratory species. Attempts to design conservation landscapes across large regions based on models of distributions in a single season or a small region may not fully reflect the limiting factors that are driving population declines.

Our goal is to predict and explore patterns of species' occurrence and local habitat usage across broad landscapes. We use data from eBird (http://www.ebird.org), an online citizen science bird-monitoring project and environmental descriptions from continent-wide covariates linked through observation location and time. These covariates include remotely sensed habitat information from the National Land Cover Database and vegetation phenology from MODIS. We model species occurrence with the SpatioTemporal Exploratory Model (STEM), an ensemble model designed to adapt to non-stationary spatiotemporal processes. This is accomplished by creating a large ensemble of local models, each restricted to a local spatial and temporal region. Within each region a user specified predictive model associates the predictors with the response. Patterns modeled locally "scale up" via ensemble averaging to larger scales.

Here we analyze eBird data to study broad-scale movements of bird populations throughout the year. We use STEM built with decision trees to adapt to a wide variety avian migration patterns without requiring a detailed understanding of the underlying dynamic local processes. We demonstrate how eBird data are capable of resolving the changing distributions of birds through their migrations. Then we illustrate how seasonal variation in habitat association can be identified and explored. These tools provide valuable information for generating hypotheses and making inference about the processes driving dynamic species distributional patterns.

Keywords: multi-scale, spatiotemporal, citizen science, non-stationary.

doi:10.1016/j.proenv.2011.07.010

1. Introduction

More than ever, research to identify the environmental drivers that shape species' distributions is needed to manage and conserve earth's natural systems. However, obtaining this knowledge is challenging because: 1) species' distributions vary dramatically through time and space across a range of spatial and temporal scales, 2) detailed species observation data are difficult to collect and organize across sufficiently large scales, and 3) conventional analytical methods have not been effective for facilitating spatiotemporal pattern discovery with such sparse, noisy data and highly variable ecological signals. The goal of our research program has been to advance data intensive ecology [1] to meet these challenges and improve our understanding of the broad-scale dynamics of continent-scale bird migrations.

By associating environmental inputs with observed patterns of bird occurrence, predictive models provide a convenient framework to harness available data for predicting species' distributions and exploring predictor effects. In this paper we use the SpatioTemporal Exploratory Model (STEM) [2], a recently developed model designed to adapt to non-stationary spatiotemporal processes, to analyze species occurrence distributions from data collected in eBird (http://www.ebird.org), an online citizen science bird-monitoring project. We show how STEM captures inter- and intra- annual changes in species' occurrence distributions. At the level of species-habitat associations, ecologists need to understand which habitats types are most strongly associated with species' occurrence, and how these associations change through time and across space. We also show STEM output can be studied to discover the spatial structuring of species-habitat associations. With this information ecologists will be better able to identify, prioritize and coordinate conservation actions across broad landscapes.

In the Section 2 we describe the bird occurrence and environmental data. Section 3 introduces the STEM species distribution model and the spatiotemporally indexed predictor importance measures used to study species-habitat associations. Results are presented in Section 4 followed by a brief conclusion in Section 5.

2. Data

The bird observation data comes from the citizen science project, eBird [3, 4]. eBird is unique among broad-scale bird monitoring projects in that it collects observations made throughout the year. Participants follow a protocol where time, location, and counts of birds are all reported in a standardized manner. By asking participants to indicate when they have recorded all the species detected, we can assume that species with no detections convey absence information for that observation. For this analysis we only used data where the participants recorded all detections and provided additional information on search effort. Together, the reports of absence and effort information add valuable information allowing the analytical control of variable detection rates when inferring absences. For each species studied, we analyzed presence-absence data from observations collected during the six-year period 2004-2009 within the conterminous U.S. There are 622,124 observations reported from 107,295 unique locations within this area. Tied to each observation are covariates that quantify search effort and describe the surrounding environment. Four effort variables are included in the analysis to account for variation in detection rates (the hours spent searching, the kilometer length of the transect, observation time of day, and the number of people in the search party). The day of the year is included to capture day-to-day level variation. To account for habitat-selectivity each observation is linked to remote-sensing data from the U.S. 2001 National Land Cover Database where landcover is classified into one of 16 classes with 30m pixels. This information was summarized as the percent coverage and spatial configuration for each vegetation class

within a 1.5km pixel centered at the observation location using FRAGSTATS [5]. Elevation, Hydrography (U.S. National GAP program) and Normalized Difference Vegetative Index [6] and climate information (mean monthly snow depth, amount of precipitation, mean temperature, minimum temperature and maximum temperature with 4km pixel) from Climate Atlas of the US (1961–1990) were also included. To account for additional anthropogenic effects we used human population density estimates from the U.S. Census Bureau 2000 census block-level summaries. The species distribution models presented here use between 20 and 70 covariates at a time depending on the specific goals of the analysis.

3. Methods

In this section we briefly describe STEM and how it is used to analyze spatial structure in species-habitat associations.

3.1. The SpatioTemporal Exploratory Model

STEM [2] is an ensemble model designed to adapt to non-stationary spatiotemporal processes. This is achieved by creating a randomized ensemble of overlapping local models, each applied across a restricted geographic and temporal extent or *stixel*. A user-specified predictive model accounts for local variation as a function of local predictor values. Predictions are made for explicit location-time pairs by taking the mean across all of the overlapping local models that include that location-time. Thus, local patterns are allowed to "scale up" via ensemble averaging to larger scales. This combines the bias-reducing properties of local models (e.g. decision trees, [6]) with the variance-reducing properties of randomized ensembles (e.g. bagging, [7]).

Formally, let $y_i, i = 1,...,N$ be a set of responses each associated with p predictors $\mathbf{x}_i = [x_{1,i} \; \ldots \; x_{p,i}]$. It is assumed that each observation, y_i, conditioned on $\mathbf{x}_i$, arises as a realization from some true but unknown function, $F^*(\mathbf{x}_i)$ that maps $\mathbf{x}_i$ to y_i. The STEM ensemble is a discrete mixture model

$$F(\mathbf{x},s,t) = n^{-1}(s,t)\sum_{i=1}^{M} f_i(\mathbf{x},s,t)I((s,t) \in \theta_i),$$

where M is the size of ensemble and each base model $f_i(\mathbf{x},s,t)$ is a function of the predictors $\mathbf{x}_i$ indexed at location s and time t and defined on the stixel support set, θ_i. $I((s,t) \in \theta_i)$ is the indicator function, taking value 1 when location and time are within support set θ_i, and zero otherwise and the function $n(s,t) = \sum I(s,t) \in \theta_i$ calculates the number of ensemble models supporting the prediction at (s,t). The ensemble used here consists of a large set of equal sized, randomly located stixels with substantial overlap to facilitate ensemble averaging (more details can be found in [2]). We have found that we can generate detailed, continent-wide predictions with a stixel size of 12 degrees longitude by 9 degrees latitude by 40 days. Decision trees (DTs) are used as base models because they have several features that make them a good choice for exploratory analysis; they automatically identify the most important predictors and the function form of their effects, including interactions [6].

3.2. Species' Distribution Estimates

For each species we calculate one daily occurrence map per week for all 52 weeks in 2009. Each weekly distribution surface is estimated with 130,769 locations selected from a geographically stratified random design with 15 locations sampled uniformly from each ~30 km pixel across a regular grid. The finest spatial resolution of each prediction is determined by the spatial resolution of the predictors, 1.5 to

4km. Thus, to make inferences across larger spatial areas areal expectations can be approximated as Monte Carlo averages of all the occurrence predictions that fall within the desired area. Variation in detectability associated with observation effort is controlled by assuming that all effort predictors (search time, transect length, time of day, and number of observers) were constant and additively associated with the true occurrence probability. Thus, the maps show the estimated probability that a single eBird participant will detect the species on a search from 7-8AM while traveling 1km on the given day of 2009, averaged across that pixel. We use a data-folding procedure to reduce bias in geographical areas with the lowest density of observations and to provide estimates of prediction variability. For each of ten data-folds, a random subset of training data is selected without replacement from the unique locations. To improve computational efficiency, we subsample 80% the training data. Estimated occurrence rates are computed as fold-averages.

3.3. SpatioTemporal Estimates of Predictor Importance

To study how habitat use changes across a species' range and throughout its' annual cycle we analyze spatiotemporally explicit measures of Predictor importance (PI). Following [7], the importance of the *k*-th DT predictor is computed as the sum of the empirical improvement in the DT splitting criterion due to this predictor. The PI over an ensemble of trees is defined as the sum of PI values for each tree in the ensemble [9]. To make the ensemble PI measure spatiotemporally explicit, we average over only those base models with support within the specified region and season. Let, $PI^k(S,T)$ be the importance of predictor *k* within region *S* and time interval *T,* then $PI^k(S,T) = |E|^{-1} \sum PI^{i,k}$ where $PI^{i,k}$ is the variable importance of the *k-th* predictor in the *i-th* base model where the sum is taken over all base models with support in region *S* and time interval *T*, and $|E|$ is the number of base models in the sum. The DT fitting process sequentially selects predictors and then fit their effects, so $PI^k(S,T)$ will be zero when the predictor is not selected and increasing values indicating both the frequency of use and the strength of impact the predictor has on model predictions. Relative predictor importance is often used to study differences between different regions and/or seasons because importance increases with species' prevalence. Note, $PI^k(S,T)$ does not covey information about the functional form or directionality of predictor effect. Partial dependence functions [9] can be used to measure the direction and functional form of association.

4. Results

Figure 1 shows the spring migration of Wood Thrush in the eastern U.S., demonstrating the ability to model within-year distributional dynamics with eBird data. Detailed predictions reveal both broad-scale patterns and fine-resolution detail. An advantage of this methodology is its adaptability without requiring detailed information about the underlying dynamic processes. The same methodology, using the same model parameters and initializations, has produced high quality distribution estimates for many species with a wide variety of spatiotemporal patterns. Monthly predictive performance during periods of residency in the continental U.S. achieved AUC scores above 0.85 and Kappa values above 0.5 for many species (unpublished). Animated distributions for more species can be seen at http://ebird.org/content/ebird/about/occurrence-maps/occurrence-maps.

In addition to understanding species' distributions, ecologists also need to understand the biological processes that give rise to distributional patterns. For example, bird species with very broad geographical distributions must, necessarily, adapt to differences in local habitat availability across their range. To test the ability of this analysis to detect these differences, we analyzed the distribution of Northern Cardinal*,* a

well-studied, non-migratory species with a broad distribution. Figure 2 shows boxplots of the relative variable importance of populations living in the southwestern U.S. state of Arizona and in the northeastern U.S. state of New York. The differences in habitat use shown in this plot match known differences and demonstrate the ability of the analysis to adapt to spatial variation in local-scale species-habitat associations that commonly arise from multi-scale ecological processes.

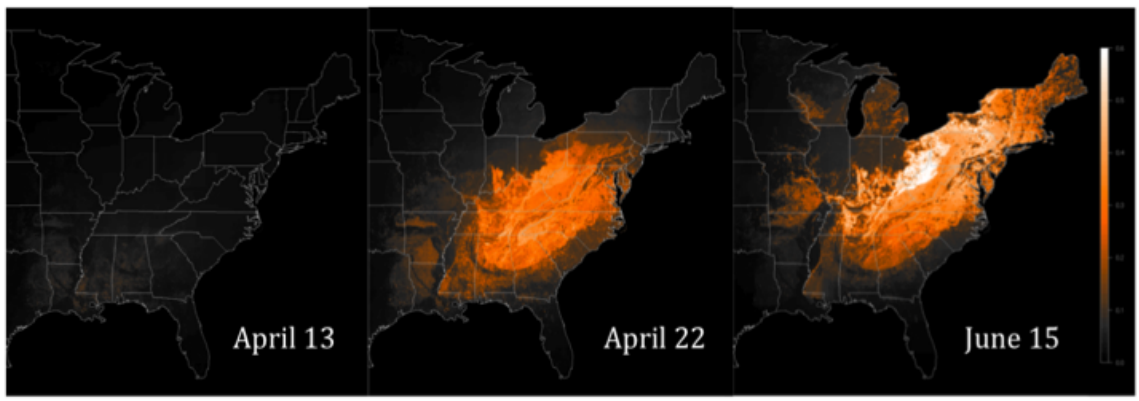

Fig. 1. Spring Migration of Wood Thrush. The predicted probability of occurrence for Wood Thrush (Hylocichla mustelina) is shown during the spring migration in 2009. The Wood Thrush population crosses the Gulf of Mexico and begins to make landfall in the South Eastern U.S. by April 13. By April 22, the population has begun its northward expansion and by June 15 it has completely filled in its breeding distribution with notable concentrations in the deciduous forests.

5. Conclusions

In this paper we demonstrated a highly automated method for modeling non-stationary spatiotemporal processes using broad scale biodiversity data collected by citizen scientists. Using this approach it is possible to detect and describe changes in observed distributions both through time and space. Moreover, the method provides a means to conduct exploratory inference using spatiotemporally explicit measures of variable importance. These variable importance statistics provide information about how, when, and where ecological processes change. We believe that these methods will be increasingly important for broad-scale conservation applications and for other broad-scale environmental applications characterized by multi-scale processes.

Acknowledgements

We thank the thousands of eBird participants for providing the data used in this analysis, the THM and staff in the Information Sciences unit at the Cornell Laboratory of Ornithology for their work in managing these data. The following organizations and programs provided support this project: the Leon Levy Foundation, the Wolf Creek Foundation, and the National Science Foundation through DataONE (0830944), the Institute for Computational Sustainability (0832782), research grant (1017793), and TeraGrid computing resources provided under grant number [TG- DEB100009].

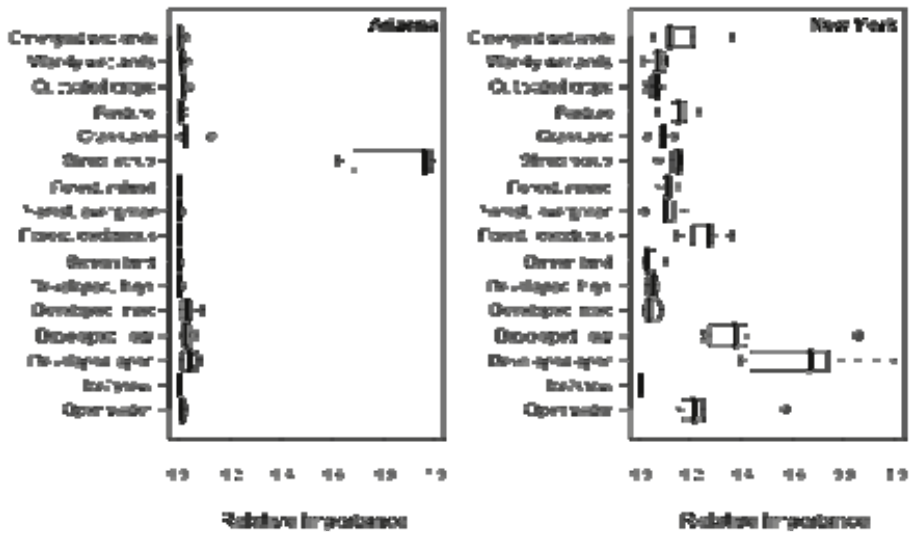

Fig. 2: Relative variable importance for Northern Cardinal (*Cardinalis cardnalis*) in Arizona and New York. Variation in boxes is the fold-level variation. The most important habitat predictor in Arizona is "shrub/scrub" while the most important habitat predictors in New York are "developed open" and "developed low", categories describing parklands and suburban areas.

References

[1] Kelling S, Hochachka WM, Fink D, Riedewald M, Caruana R, Ballard G, and Hooker G. Data-intensive Science: A New Paradigm for Biodiversity Studies. *BioScience* 2009; 59: 613-620.

[2] Fink D, Hochachka WM, Zuckerberg B, Winkler DW, Shaby B, Munson MA, Hooker GJ, Riedewald M, Sheldon D, Kelling S. Spatiotemporal Exploratory models for Large-scale Survey Data. *Ecological Applications* 2010; 20(8): 2131-2147.

[3] Sullivan, B., Wood, C. Iliff, M. J. Bonney, R. E. Fink, D. and Kelling, S. 2009. eBird: A Citizen-based Bird Observation Network in the Biological Sciences: Biological Conservation 142, 2282–2292.

[4] Munson, M. A., K. Webb, Sheldon, D. , Fink, D., Hochachka, W.M., Iliff, M., Riedewald, M., Sorokina, D., Sullivan, B., Wood, C., and Kelling, S. 2010. The eBird Reference Dataset 2.0. (http://www.avianknowledge.net/content/features/archive/eBird_Ref)."

[5] McGarigal, K., S. A. Cushman, M. C. Neel, and E. Ene. FRAGSTATS: Spatial Pattern Analysis Program for Categorical Maps. University of Massachusetts, Amherst. Version 3. Available from: http://www.umass.edu/landeco/research/fragstats/fragstats.html.

[6] Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC). 2009. MODIS subsetted land products, Collection 5. Available on-line [http://daac.ornl.gov/MODIS/modis.html] from ORNL DAAC, Oak Ridge, Tennessee, U.S.A. Accessed November 20, 2009.

[7] Breiman, L, et al., *Classification and regression trees*. New York, New York: Chapman & Hall; 1984.

[8] Breiman, L., Bagging predictors. *Machine Learning* 1996; 24: 123-140.

[9] Hastie T, Tibshirani R, and Friedman J. *The elements of statistical learning: data mining, inference, and prediction, 2nd edition*. Springer Verlag, New York, New York, USA, 2009.